Rev. Téc. Ing. Univ. Zulia, Vol. 15, No. 2, 87 - 94, 1992

Estudio de la distribución espacial de la precipitación mensual en la cuenca del Lago de Maracaibo

S. de Bautista y J.J. Villasmil.

División de Postgrado Facultad de Ingeniería Universidad del Zulia, Maracaibo, Venezuela

Resumen

En el presente trabajo se estudian los aspectos teóricos del Análisis en Componentes Principales desde el punto de vista analítico, algebráico, y geométrico y su representación gráfica. Se discuten los criterios de selección de los Componentes Principales significativos.

Se aplica el método a un problema del campo de la Hidrología, el estudio de la distribución espacial de la precipitación mensual en la Cuenca del Lago de Maracaibo a partir del análisis del comportamiento de los Componentes Principales calculados de la Matriz de Correlación de los datos mensuales de precipitación medidos en las estaciones ubicadas en las diversas sub-cuencas del Lago.

Se concluyó que las variables: altitud, orientación de las sub-cuencas con respecto a la dirección de los vientos predominantes, ubicación en la Cuenca del Lago, forma de la cuenca inciden sobre la distribución espacial de la precipitación mensual.

Spatial distribution of monthly precipitation in Lake of Maracaibo basin

ABSTRACT

In this paper we present the theoretical aspects of the method of Principal Components Analysis, from a analitical, algebraic and geometric viewpoint and also their graphic representation. We discuss the selection criteria of the first significant Principal Components.

As an application of the method in Hidrology it was studied the spatial distribution of the monthly precipitation in the Lake of Maracaibo's basin. With the analysis of the behaviour of the Principal Components calculated from the Correlation Matrix of the monthly precipitation gaged in different stations ubicated in the sub-basins of the Lake it was concluded : altitude, orientation of the sub-basins respect the direction of the predominant winds, the form of the basins, incides on the statial distribution of the monthly precipitation.

Introducción

El Análisis de Componentes Principales (A.C.P.) corresponde a un método de análisis multivariado (conjunto de P variables y N observaciones) que tiene por objeto examinar formas simplificadas de representar el complejo que se estudia, transformando un conjunto de variables interdependientes a independientes, o reducir la dimensionalidad del complejo.

El A.C.P. transforma el grupo de P variables correlacionadas en un nuevo grupo de variables no correlacionadas (Componentes Principales) que son combinaciones lineales de las variables originales y están derivadas en orden decreciente de importancia, de modo que la primera C.P. va a explicar tanto como sea posible la variación en los datos originales. Si los "q" Primeros C.P. dan respuesta a la variación en los datos originales,

se puede decir que la dimensionalidad del problema es menor que "P".

1.I. Derivación de los C.P. en forma matricial

Sea X una variable aleatoria con dimensión "P", media $\overline{X}$ y matriz de covarianza C.

$X^T = X_1,.....,X_p$ (matriz de vectores columnas)

Se encontrará una variable Z (Componentes Principales) con dimensión "P", tal que:

$Z = A^T.\ X$ donde A^T es un vector de constantes

Se calcula la primera Componente Principal (Z_1) escogiendo un vector a_1, t.q.Z_1 exprese la mayor varianza posible.

$$Z_1 = a_1{}^T X \qquad (1)$$

Var $(Z_1) = a_1.^T.C.a_1$ con C = matriz de covarianza (2)

maximizando (2) sujeto a $a_1T.a = 1$, se obtiene:

$$(C - \lambda I)\ a_1 = 0$$

Para una solución a_1 distinta a la trivial, debe escogerse λ, tal que $|C - \lambda I| = 0$ ... (3)

Las raices de la ecuación del determinante (3) son los valores propios de C. La ecuación (3) tendra "P" raices no negativas. $\lambda 1 \geq \lambda 2 \geq ..> \lambda$ P

Determinando la varianza del Primer Componente Principal queda:

$$\text{var}\ (Z_1) = \lambda_1.\quad a_1T.a = 1\ ..\qquad (4)$$

Si se quiere maximizar la varianza se deberá escoger el mayor valor propio, λ_1.

De esta manera, a_1 sera el vector propio de C correspondiente al mayor valor propio.

Geométricamente consiste en buscar para el Primer Componente Principal un eje tal que la suma de los cuadrados de las distancias de los "N" puntos al eje sea minima. Esta distancia deberá ser perpendicular al eje, por lo tanto se impone una condición de normalización,, $a_1{}^T.$ a=l.

En resumen $Z_1 = a_1{}^T.\ X$ constituirá el primer eje tal que el cuadrado de las distancias de los N puntos a dicho eje sea minimo. Se demuestra que este eje pasa por el centro de graveded de los N puntos.

El segundo Componente Principal $Z_2 = a_2{}^T X$ se obtiene por una extensión del argumento anterior, teniendo que $a_2\ Ta_2 = 1$, e imponiendo una segunda condición que establece que Z_2 no esté correlacionada con Z_1.

$$\text{cov}\ (Z_2, Z_1) = a_2TCa_1 = 0 \qquad (5)$$

Maximizando la varianza de Z_2, sujeta a $a_2{}^Ta_1 = 0$ y $a_2{}^Ta_2 = 1$ se obtiene la expresión $(C - \lambda I)\ a_2 = 0$ donde será el segundo mayor valor propio de C y a_2 su correspondiente vector propio.

Desde el punto de vista geométrico, se determina un segundo eje, proyectando N puntos a un hiperplano ortogonal al Primer plano, tal que la suma de los cuadrados de las distancias de los puntos al eje sea minimo.

El procedimiento de cálculo se repite "P" veces. La suma de los cuadrados de las distancias de todos los puntos a su centro de graveded puede ser expresada como la suma de los valores propios o si esta suma es S, se puede decir que el Primer Componente Principal "explica" una proporción λ_1/S de la variación total, los primeros dos Componentes Principales aportarán una proporción $(\lambda_1 + \lambda_2)/S$, y asi sucesivamente.

En particular, si ciertos λ's poseen valores muy pequeños, dando como resultado $q \leq P$ valores significativos se dice que la variación ocurre en un espacio de menor dimension, q.

Si en lugar de trabajar con la matriz de datos X, con media $\overline{X}$ y desviación σX, se forma una nueva matriz X con datos normalizados, con media cero y desviación standard unitaria, se puede realizar el mismo ACP pero sustituyendo la matriz de covarianza C por la matriz de correlación R, ya que

$$R = \frac{1}{N}\ X^T\ X \quad [X_i] = \frac{X_{ij} - \overline{X}_j}{\sigma Xj} \quad \text{donde} \qquad (6)$$

los valores propios obtenidos y sus vectores propios corresponden a la matriz R. En este caso

$$\sum_{i=1}^{P} \lambda i = P \qquad (7)$$

Correlación entre una variable y un C.P.

Es importante conocer la correlación entre cualquier variable X_j y cualquier Componente Principal Z_k [1,2] para establecer el porcentaje de información que se pierde en cada vaxiable X, al trabajar con esa C.P. Se deduce que:

$r_{jk} = a_{kj}\sqrt{\lambda k}$..... r_{jk} = coef. de correlación (8)

Correlación entre una variable y más de un C.P. [1,2]

Como los C.P. no están correlacionados, se tiene que:

$$R^2(X_j,Z_1,Z_2,\ldots) = r^2(X_j,Z_1) + r^2(X_j,Z_2)+\ldots \qquad (9)$$

Representación gráfica de los componentes principales

El uso de la representación gráfica facilita la interpretación de los resultados y presentar de manera visual los resultados que permiten conocer la estructura de los datos.

Representación Cartográfica de cada serie a_{ij}. (j = 1,.....P): [1,2] sobre un mapa donde se encuentran ubicadas las estaciones se trazan las lineas de igual coeficiente a_{ij} (isocosenos directores) correspondientes a un mismo C.P. Se facilita la interpretación fisica de los C.P.

Elipses de Proximidad: [1,2] Se construyen a partir de variables **a** las cuales interesa estudiar su grado de proximidad. En el caso de estudios de redes de estaciones, permite observar estaciones redundantes o faltantes. Se pueden trazar para los pares de valores r_{j2}, r_{j3} y eventualmente para r_{j4}, r_{j5}. Parte de la suposición que los coeficientes de correlación poseen una función de densidad de repartición gaussiana-bidimensional. O sea, si

$$F(Y,T) = \frac{1}{2S_Y \cdot S_T \sqrt{1-r^2}}\, e^{-E^2/2} \qquad (10)$$

r_{j2} = Y; r_{j3} = T con j = 1,...., m con (m- t,) <P serán el grupo de variables agrupadas.

donde

$$E^2 = \frac{1}{1-r^2}\left[\left(\frac{Y-\overline{Y}}{SY}\right)^2 - 2r\left(\frac{Y-\overline{Y}}{SY}\right)\left(\frac{T-\overline{T}}{ST}\right) + \left(\frac{T-T}{ST}\right)^2\right] \qquad (11)$$

$\overline{T},\overline{Y}$ = medias: SY, ST = desv. tipicas

r = coeficiente de correlación entre Y,T

Se efectúa un cambio de ejes:

$u = (Y-\overline{Y})\cos\theta + (T-\overline{T})\,\mathrm{sen}\theta$

(12)

$v = (Y-\overline{Y})\,\mathrm{sen}\theta + (T-\overline{T})\cos\theta$

Se escoge θ tal que u y v sean independientes y se obtiene una ecuación:

$$X^2 = \frac{u^2}{Su^2} + \frac{V^2}{Sv^2}$$

donde:χ : distribución chi-cuadrado (13)

A partir de la cual se puede construir las elipses de proximidad al % .

$$\int_0^{\chi^2} \frac{1}{2} e^{\chi^2/2} \cdot d\chi = P$$

2. Criterios para la escogencia de los "q" primeros componentes pruebas de hipotesis: La mayoria de las pruebas de hipótesis desarrolladas parten de la presencia de una distribución multinormal y se aplican a matrices de varianza. En el trabajo se aplicó la prueba de hipótesis que los (p-q) valores propios de Σ son iguales [3] (MARDIA, 1979) con la modificación sugerida por BARTLETT (1951) [3] (MARDIA, 1979) para el caso de matrices de correlación.

Se calcula el valor

(n-1) (p-q) log (ao/go)

ao: media aritmética de los valores repetidos

go: media geométrica de los valores repetidos.

Se compara con el valor tabulado

X^2 (1/2.(p-q + 2) (p-q-1)

Si el valor calculado es mayor que el valor tabulado, se rechaza la hipótesis, es decir los (p-q) últimos componentes no son iguales.

Al aplicar la prueba a los valores propios calculados, a partir de la Matriz de Correlación, se observó que es una prueba muy susceptible al tamaño de la muestra.

Criterios prácticos para la escogencia del número adecuado de componentes principales.

-CATTELL (1966) [3] (MARDIA, 1979) planteó la construcción del gráfico λi vs i.

-Incluir el número de Componentes que expliquen el 90% de la variación total. [3] (MARDIA , 1979)

KAISER [3] (MARDIA, 1979) sugiere excluir aquellos componentes cuyos valores propios poseen su valor menor que el valor promedio de los valores propios de la matriz.

No escoger más de cinco Componentes Principales.

3. Planteamiento del problema

OBJETIVO: Analizar la distribución espacial de la precipitación mensual medida en algunas sub-cuencas del Lago de Maracaibo.

METODOLOGIA: Recolección de la Información Básica: a) Delimitación, b) Obtención y Procesamiento de la información Pluviométrica[4]. -Aplicación del Método de C P.: a) Cálculo de -Matriz de Correlación, -Valores y Vectores Propios. (Fig. 1); escogencia de los "q" C.P. d) Análisis de Resultados (Tabla No. 1, e) Conclusiones.

SISTEMAS ESTUDIADOS	No. ESTACIONES
Palmar-Apón	13
Apón-Yasa	11
Santa Ana	10
Escalante	13
Motatán	17
	64

Se analizó la precipitación mensual en el período 1973,1983, o sea 11 años, 132 observaciones por cada estación.

ESQUEMA GENERAL DEL PROGRAMA PRINCIPAL

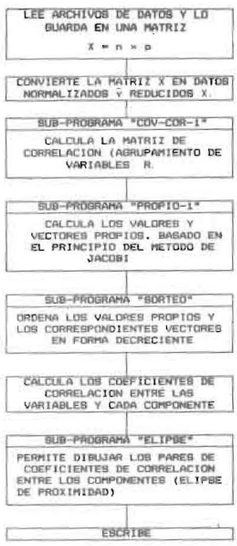

Figura No. 1

Se presentaron las gráficas de la aplicación del Método al Sistema Motatón: Fig. 2: Representación Cartográfica de los $a1_j$, Fig.3: Representación Cartográfica de los $a2_j$, Fig. 4: Representación Cartográfica de los $a3_j$, Fig.6: Elipses de proximidad ($\rho 2_j$ vs $\rho 5_j$) Fig. 7. Elipse de Proximidad $\rho 4_j$ vs $\rho 5_j$. Fig. 5 Gráfico λ_i vs i. Tabla de Resultados de los 5 sistemas estudiados. (TABLA No. 1)

4. Conclusiones referentes al método

El análisis en Componentes Principales permite analizar la estructura de correlación de un conjunto de variables que se han medido para estudiar un fenómeno. Este análisis

TABLA No. 1

SISTEMAS	No. ESTACIONES	FORMAS DE LAS CUENCAS	ORIENTACION CON RESPECTO AL CENTRO DEL LAGO	RANGO DE ALTITUDES m.s.n.m.	VALORES PROPIOS ACUMULADOS (%)
Palmar-Apón	13	Ovoide	NW	82-242	1° 73,36 2° 80,78 3° 84,01 4° 87,46 5° 90,30
Apón-Tasa	11	Ovoide	WNW	77-1524	1° 74,16 2° 81,97 3° 87,94 4° 90,80 5° 93,10
Santa Ana	10	Ovoide	WSW	38-1524	1° 74,73 2° 81,73 3° 87,86 4° 91,13
Escalante	13	Ovoide	S	2-2114	1° 59,28 2° 66,76 3° 72,53 4° 77,77 5° 81,97
Motatán	17	Achatada	ESE		1° 56,64 2° 71,13 3° 75,32 4° 79,07 5° 82,34

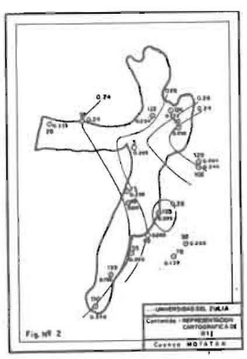

Figura 2

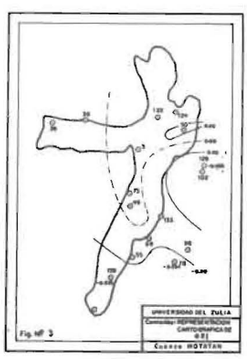

Figura 3

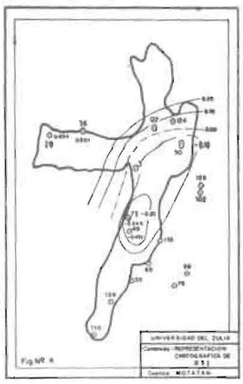

Figura 4

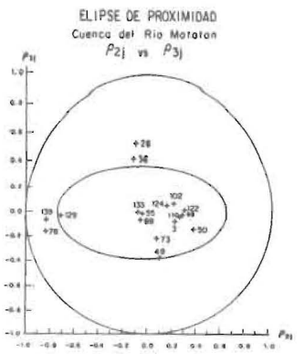

Figura 6

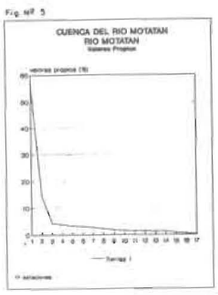

Figura 5

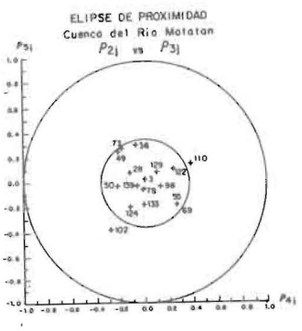

Figura 7

permite a su vez definer los factores no medidos que agrupan o separan estas variables. En el caso de la precipitación se observó que la altitud de las estaciones medidoras de lluvias, su cercanía a la fuente de humedad, la forma de la cuenca y su orientación con respecto a los vientos constituían factores que agrupaban o separaban a los valores de precipitación mensual medidas en las estaciones.

El uso de las distintas formas de representación gráfica de los componentes se considera imprescindible, sobre todo en los casos en que se maneja un alto número de variables.

Si las variables miden unidades distintas, deberá trabajarse directamente con la matriz de correlación.

Para la determinación de los "q" Primeros Componentes significativos, se sugiere aplicar los criterios prácticos, y no escoger más de 5 componentes, ya que es muy difícil obtener el significado físico más allá de ese valor. El problema de las pruebas de hipótesis es que son sensibles al tamaño de la muestra.

5. Conclusiones referentes a la aplicación

TABLA No. 2

SISTEMA	VARIABLES	C.P.	% VARIABILIDAD
Palmar-Apón	13	5	90.30
Apón-Yasa	11	4	90.80
Santa Ana	10	4	91.13
Escalante	13	5	81.97
Motatán	17	5	82.34

Estos métodos facilitan la interpretación física de los componentes y el estudio de los agrupamientos de las variables.

Si las variables a estudiar miden las mismas unidades, es decir, en todas se mide altura, o lluvia, es necesario examinar la matriz de varianza-covarianza. En el caso de que los valores de las varianzas difieren mucho, deberán calcularse los Componentes Principales a partir de la matriz de correlación.

En todos los casos se redujo a la dimensión del problema a 4 ó 5 Componentes Principales.

En los sistemas estudiados el Primer Componente (Z_1) constituye la precipitación media mensual en el área. Este valor aporta los siguientes porcentajes de variabilidad en cada caso:

TABLA No. 3

SISTEMA	% VARIABILIDAD DE Z_1	UBICACION	FORMA
Palmar-Apón	73.76	NW	Ovoide
Apón-Yasa	74.16	NW	Ovoide
Santa Ana	74.73	W	Ovoide
Escalante	59.28	S	Ovoide
Motatán	56.64	E	Achatada

Este resultado es importante dado que permite cuantificar el porcentaje de variabilidad que aporta la media en el área.

Los Componentes Principales 2 y 3 reflejan la incidencia de la altura en la variabilidad total.

Se pudo observar la variabilidad introducida por las estaciones ubicadas en zonas altas o zonas bajas, o los efectos opuestos entre estaciones ubicadas a diferentes alturas.

Los Componentes Principales Cuatro y Cinco aportan variabilidades muy pequeñas provenientes de algunas estaciones que en algunos casos por su ubicación ya sea muy cercanas o muy altas, introducen porcentajes de variabilidad que se reflejan en dichos componentes.

En las cuencas ubicadas en la Costa Occidental el Primer Componente aporta más de un 70% en las cuencas ubicadas al sur y en la parte oriental no llega al 60%.

A partir del estudio de la interpretación física de los Componentes Principales se puede concluir que las variables que deben tomarse en cuenta para la distribución de estaciones medidoras de precipitación, en una cuenca como la del Lago de Maracaibo (fuente de humedad), son: Precipitación Media mensual, Altura sobre el nivel del mar, Distancia de la fuente de humedad, Forma de la cuenca, Dirección y sentido de los vientos predominantes de la zona.

TABLA DE NOTACION

X^T = matriz de datos $\begin{bmatrix} (X11-X1)\ (X27-X2)\ldots\ (XP1-Xp) \\ (X12-X1)\ (X22-X2)\ldots(XP2-Xp) \\ . \\ . \\ . \\ (X1n-X1)\ (X2n-X2)\ldots\ (Xpn-Xp) \end{bmatrix}$

X1,X2, ... XP valores medios de los "p" variables.

Z = matriz de componentes principales = $A^T.X$

A^T = matriz de vectores propios (pxp) = $\begin{bmatrix} A11\ A22\ldots\ Alp \\ A21\ A22\ldots..A2p \\ . \\ . \\ Apl\ \ Ap2\ldots\ldots App \end{bmatrix}$

(Aij) = vector propio o cosenos directores correspondientes al i- ésimo componente principal y la j-ésima variable.

C = matriz de covarianza de la matriz de datos X

λi= valor propio correspondiente al i-ésimo componente principal.

r_{jk} = coeficiente correlación entre la j-ésima variable y el k-ésimo componente principal.

Referencias Bibliográficas

1) BOIS, P.: "Apuntes del CIDIAT". Mérida, Venezuela, 1981.

2) DUBAND, D.: "Notas del Curso Hydrologic Statistique Approfondie". Ecole Nationale Superieure D'Hydraulique. Grenoble. Francia. 1982.

3) MARDIA, K. V.: "Multivariate Analysis. Academic Press. Chapter 8. 1979.

4) RINCON, B.: "Modelo computarizado para el Manejo de Información a nivel Mensual". Trabajo de Ascenso. Facultad de ingenieria. Universidad del Zulia. Maracaibo. 1986.

Recibido: 27 de Abril de 1990
En forma revisada: 02 de Octubre de 1991